Comparison of endogenous retroviral RNA profiles from blood cells and plasma, between nonagenarians and young controls

Master's Thesis

Arttu Autio
Bioinformatics Master's Degree Programme
Faculty of Medicine and Life Sciences
University of Tampere
2018

Acknowledgements

This Master's Thesis project was done in the Faculty of Medicine and Life Sciences at the University of Tampere, Finland. I am very grateful to Professor Mikko Hurme for granting me this opportunity to complete my Master's Thesis with a thesis project that aligns with my own research interests. I have been lucky to get to work with also Tapio Nevalainen, Saara Marttila and Binisha Hamal Mishra. I have enjoyed our many discussions, both work-related and otherwise. Professor Mikko Hurme and Tapio Nevalainen both have gone above and beyond to instruct me on not just the topics of this work, but also on how one should conduct research in general. The teachings of Professor Matti Nykter, Juha Kesseli and Kirsi Granberg form the foundation of my bioinformatics knowledge, and I hope that I may one day master all the methods they have introduced me to. I am also grateful for Marja Jylhä for spearheading the research on the very eldest and being instrumental in the generation of the datasets that were studied in this work. Finally, I would like to thank the people closest to me for being patient, as I was engrossed in finishing this work.

Tampere, December 5, 2018
Arttu Autio

Master's Thesis

Location:	University of Tampere, Faculty of Medicine and Life Sciences
Author:	Autio, Arttu Oskari Juhani
Title:	Comparison of endogenous retroviral RNA profiles from blood cells and plasma, between nonagenarians and young controls
Pages:	50
Supervisor:	Professor Mikko Hurme
Reviewers:	Professors Matti Nykter and Mikko Hurme
Date:	December 2018

Abstract

Background and Aims: An estimated 8% of the human genome consists of integrated virus genomes that have been left over from past retroviral infections. These human endogenous retroviruses (HERV) have been ravaged by mutational decay, often over millions of years, and most have been rendered inactive. However, some still contain open reading frames and may even code for functional protein products. Upregulation of these retroelements has been observed in certain disease states and in aging mice.

Materials and Methods: PBMC and plasma samples were previously obtained from nonagenarians (n=7, age 90) and young controls (n=7, age 26-32, median age 28). RNA-sequencing of the samples was performed and the expression of endogenous human genes and HERV proviruses of the families HERV-K (HML-2) and HERV-W was quantified. This thesis work focused on the characterization and comparison of proviral expression values between the biological sources and across age groups.

Results: Proviral expression differed greatly between PBMCs and plasma. Age-associated significant differential expression of three HERV-K (HML-2) and one HERV-W provirus was found in PBMCs. Furthermore, hierarchical clustering of samples indicated aging-associated differences in proviral expression patterns. No age-associated differences in proviral expression could be found in plasma samples.

Conclusion: There are age-associated differences in the expression of HERV-K (HML-2) and HERV-W in PBMCs. Between sample differences in proviral expressions in plasma were too great to determine differences between age groups. The differing proviral expression between PBMC and plasma supports the hypothesis of enriched proviral cargo of extracellular vesicles (EV). It was not possible to determine with confidence the origin of the EVs in the plasma samples. Investigation of proviral expression using the same approach with larger datasets of RNA-sequencing data, potentially from external databases, could offer further insights.

Pro Gradu -tutkielma

Paikka:	Tampereen yliopisto, Lääketieteen ja biotieteiden tiedekunta
Tekijä:	Autio, Arttu Oskari Juhani
Otsikko:	Verisolujen ja plasman RNA:n HERV-profiilien vertailu yhdeksänkymmenvuotiaiden ja nuorten verrokkien välillä
Sivumäärä:	50
Ohjaaja:	Professori Mikko Hurme
Tarkastajat:	Professorit Matti Nykter ja Mikko Hurme
Aika:	Joulukuu 2018

Tiivistelmä

Tutkimuksen tausta ja tavoitteet: Arvioitu 8% ihmisen genomista koostuu endogeenisistä retroviruksista (HERV), jotka ovat menneiden retrovirusinfektioiden geneettisiä jäänteitä. Miljoonien vuosien aikana HERV provirukset ovat rappeutuneet mutaatioiden seurauksena, tehden suurimmasta osasta toimintakyvyttömiä, mutta jotkut sisältävät vielä avoimia lukukehyksiä ja jopa ohjeet toiminnallisien proteiinituotteiden tuottamiseen. Näiden retroelementtien korkeampaa ekspressiota on havaittu vanhoissa hiirissä sekä ikääntyvissä soluissa.

Materiaalit ja menetelmät: PBMC- ja plasmanäytteitä on aiemmin kerätty iäkkäiltä (n = 7, ikä 90) sekä nuorilta verrokeilta (n = 7, ikä 26-32, mediaani ikä 28). Näytteiden RNA-sekvensointi suoritettiin ja ihmisen geenien sekä HERV-K (HML-2) ja HERV-W provirusten ekspressio laskettiin. Tässä työssä keskityttiin provirusten ekspressioarvojen luonnehdintaan ja vertailuun niin PBMC- ja plasmanäytteiden kuin ikäryhmienkin välillä.

Tulokset: Provirusten ekspressio erosi suuresti PBMC- ja plasmanäytteiden välillä. PBMC-ryhmässä havaittiin kolmen HERV-K (HML-2) ja yhden HERV-W proviruksen merkitsevä differentiaalinen ekspressio ikäryhmien välillä. Lisäksi näytteiden hierarkkinen klusterointi viittasi ikääntymiseen liittyviin eroihin provirusten ekspressioprofiileissa. Plasmanäytteissä ei todettu ikäominaisia eroja provirusten ekspressiossa.

Johtopäätökset: HERV-K (HML-2) ja HERV-W perheiden provirusten ekspressiossa esiintyy ikäominaisia eroja PBMC-näytteissä. PBMC- ja plasmanäytteiden välinen differentiaali ekspressio tukee ekstrasellulaaristen vesikkelien (EV) rikastetun proviruslastin hypoteesia. Yksilölliset erot provirusten ekspressiossa plasmanäytteissä ylittävät ikäryhmien väliset erot. Ei ollut mahdollista luotettavasti määrittää EV:n alkuperää plasmanäytteistä. Samojen metodien käyttö provirusten tutkimuksessa käyttäen laajempaa aineistoa, mahdollisesti ulkoisia tietokantoja hyödyntäen, voisi tarjota lisää tietoa.

Contents

1 Introduction **1**

2 Background **3**

2.1 Endogenous retroviruses 3

2.1.1 Nomenclature 3

2.1.2 Life cycle 4

2.1.3 Biological significance 5

2.1.4 Associations with diseases 6

2.2 Proviral cargo of extracellular vesicles 7

2.3 Age-related changes in proviral expression 7

2.4 The promise of bioinformatics methods 8

3 Objectives **11**

4 Materials **12**

4.1 Sample collection 12

4.2 RNA extraction 12

4.3 Filtration of plasma samples 13

4.4 RNA sequencing 13

4.5 Quality control and read trimming 14

4.6 Alignment and annotation 14

4.7 Normalization 15

4.8 Ethics 15

5 Methods **16**

5.1 Investigation of proviral expression 16

5.1.1 Scaling of proviral expression with chosen housekeeping genes . 17

5.2 Proviral expression patterns 17

5.2.1 Principal component analysis 18

5.2.2 Hierarchical clustering 18
5.3 Inferring potential biological effects of proviruses 19
5.4 Investigation of genomic neighborhood 20

6 Results and discussion **21**
6.1 Proviral expression in blood cells compared to plasma 21
6.2 Differences in expression between age groups 27
6.2.1 Patterns of expression across proviruses 29
6.3 Origin of the extracellular vesicles 33
6.4 Inferred biological effects 34

7 Conclusion **36**

8 Future perspectives **37**

References **39**

Abbreviations

ERV	Endogenous Retrovirus
HERV	Human Endogenous Retrovirus
PBMC	Peripheral Blood Mononuclear Cell
EV	Extracellular Vesicle
LTR	Long Terminal Repeat
PCR	Polymerase Chain Reaction
PCA	Principal Component Analysis
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis

1. Introduction

Though the Human Genome Project and the ENCODE project (Encyclopedia of DNA Elements) that followed have gone a long way towards mapping the functional elements of the human genome, there still remain many regions and features that very little is known about. One of those are human endogenous retroviruses (HERVs), which comprise an estimated 8% of our genome [22]. HERVs are remnants of ancient retroviruses that have become a part of us through integration into the human genome. These endogenous retroviral elements have been ravaged by millions of years of mutational decay, rendering many of them non-functional.

In humans, HERVs have not been shown to cause disease directly [57], yet many associations have been reported. These include neurological diseases such as multiple sclerosis (MS) and schizophrenia, and many autoimmune diseases such as diabetes and systemic lupus erythematosus [10]. Changes in HERV expression have also been noticed in cancer, and they are significant enough for there to be potential for them to act as an early biomarker of some cancers [59]. In animals, some HERVs have been observed to affect biology in more extreme ways. An example of this is the intracisternal type A particle sequences (IAPs), an ERV that causes genomic damage in mice through a high transposition rate [40]. It is possible that human physiology too is affected by ERVs in yet undiscovered ways.

We have identified two promising avenues of HERV research. One is the immunodeficient state of aging. HERVs seem to behave differently in disease states and as a result of aging. ERV expression in mice has even been claimed to be "age-dependent" [58]. The other approach to HERV study concerns extracellular vesicles (EVs). Cells release EVs as a form of communication and potentially for other purposes. It has been reported that HERV mRNA and products are selectively loaded to EVs as cargo [63].

The existence of HERVs has been known since the 1960s [60], yet their nature as repetitive elements of the genome has made their study challenging. When a retrovirus integrates into a genome, there is a period of duplication during which copies of that virus spread. Therefore, many seemingly individual HERVs are derived from the same virus, causing them to share a great degree of sequence similarity. Mapping transcripts to the exact origin can thus be difficult. It can even be argued that the main barriers in HERV research are technical. The more accurately and effectively HERV expression can be measured and analyzed, the better their effects on human physiology can be determined. Newer, better hardware and software are needed to be utilized in their study. One such technology is RNA sequencing, made possible

and economically feasible by recent dramatic advances in sequencing technology.

In this work, we have utilized RNA sequencing to quantify and analyze the expression of human endogenous retroviruses in the contexts of extracellular vesicles and aging. Furthering our knowledge of HERV biology in these states could help improve our understanding of the diseases HERVs have been associated with, as well as aging as a whole. Even if no direct causality between HERVs and diseases can be demonstrated, proviral expression has already been shown to change in certain disease states and therefore there is potential for novel biomarkers.

2. Background

This section describes human endogenous retroviruses (HERVs) as well as the related topics of extracellular vesicles (EVs) and aging. Regarding HERV biology, the subtopics of nomenclature, life cycle, biological significance, as well as known associations with diseases will be covered. The promise of new bionformatics methods in the study of these topics is also considered.

2.1 Endogenous retroviruses

An estimated 8% of the human genome consists of endogenous retroviral elements [22]. These HERVs are the genetic remnants left over from past infections by retroviruses, most of which occurred millions of years ago. Each individual HERV locus in the human genome is a provirus: a virus genome that has become integrated into the DNA of a host cell. HERV proviruses are passed on vertically, from parents to offspring, as part of the genome.

2.1.1 Nomenclature

Before moving onto HERV biology, nomenclature of HERVs should be clarified. It is important to note that the classification of HERV integrations into families and subfamilies is incomplete, at times inconsistent, and often confusing. Not least due to the fact that "human endogenous retrovirus" is strictly speaking a misnomer in most cases. Very few HERVs are actually specific to humans. The vast majority of them appear to have become integrated before species divergence of humans from primates, and many date back much further. For example, some of the oldest endogenous retroviruses present in humans, HERV-S/L, have also been found in reptiles and fish. Some HERV integrations are also thought to have resulted from cross-species transmission. [14]

Traditional classification of endogenous retroviruses has been done based on sequence similarity [14]. Such similarity can indicate that proviral insertions at different loci actually originate from the same retrovirus. Heavily mutated older integrations are more differentiated [19], thereby being more difficult to classify using this approach. A now common approach to HERV classification is to separate and refer to families by the one-letter code of the amino acid that is used by transfer RNA to prime reverse transcription [57]. However, the biological effects and expression of proviruses do not necessarily correlate within HERV families derived through such classification.

Furthermore, no complete list of HERV families has been published [56].

Two notable HERV families are HERV-K (HML-2) and HERV-W. HERV-K (HML-2) is the one of the youngest HERV families, initial integration having occurred an estimated 0.2 - 2 million years ago, and is considered the most biologically active [22]. HERV-K (HML-2) is thought to be human specific, since integrations of it have only been found in humans [14]. This study uses the Subramanian et al. annotation of HERV-K (HML-2) [47], which includes 91 full-length proviral sequences. [47]

HERV-K (HML-2) has been found to be polymorphic in humans, meaning that HERV-K (HML-2) integrations are not yet fixed in the population, but rather may vary from individual to individual. Thus, there exist integrations in the population that are not present in assembled reference genomes. Proviruses of the family can be present as only solo LTRs in some individuals, while retaining a more complete, full-length genomic structure in others. [30]

HERV-W is much older in comparison, having been integrated an estimated 40 million years ago. In some ways, HERV-W has come to behave more akin to a human gene than a virus. The protein Syncytin-1 coded for by a HERV-W provirus is needed for normal placental development. This study uses the HERV-W annotation by Grandi et al. [17] of 213 full-length or near full-length elements. [17]

2.1.2 Life cycle

To understand what potential biological effects HERVs may have on their hosts, it is important to know their origins. HERV proviruses are derived from exogenous retroviruses. Similarly to RNA viruses, retroviruses store genetic material as RNA. However, retroviruses differ in that DNA intermediates are included in their life cycle. This difference is significant enough that they are classified separately from RNA viruses in the Baltimore classification of viruses [6].

To become endogenized, a retrovirus has to first infect a germ line cell that is then used in the production of offspring. That offspring must survive to reproduce and thus proliferate the provirus via its offspring. Eventually the provirus may become fixed in the population on the species level. Evolutionary selection pressures on both the genomic and organismic level affect the success of the endogenization. A retrovirus that severely reduces the chances of survival of its host will be selected out on the organismic level. Selection pressures continue to affect proviruses long after becoming fixed in the population. [10]

The base genomic structure of HERVs is similar to that of exogenous retroviruses, consisting of Gag, Pro, Pol, and Env genes and surrounded on both sides by two long terminal repeats (LTRs) [34]. Some retroviral genomes also contain code for additional

proteins and RNA [14]. However, mutational decay usually ravages HERV proviruses over time, eventually leaving most nonfunctional [22]. Many HERV's are eventually reduced to solitary LTR's. Some proviruses survive this mutational decay by instead becoming useful to the host, such as HERV-W and its useful protein product Syncytin-1. Selection pressure therefore starts working for the provirus instead of against it.

2.1.3 Biological significance

At the heart of this thesis work is the open question of the extent of biological significance that HERV proviruses hold. Despite mutational decay and epigenetic silencing rendering most HERV proviruses inactive in most contexts, there is a great deal of evidence that HERVs are not just genetic fossils. A compelling argument comes from the biological potential of the flanking LTRs. The enhancer and promoter regions of proviral LTRs are thought to be able to affect the expression of surrounding genes [26]. This means that even solitary HERV LTRs can potentially affect biology. The abundance of LTRs in the human genome could mean that their effects on human genes are widespread [10].

A more direct biological effect can result from intact proviral protein products. Some HERV-K (HML-2) proviruses are thought to contain open reading frames (ORFs) for all viral genes and to code for functional protein products [47]. Some younger HERVs are claimed to contain intact code for the production of virions and that such virus particles are capable of being released from cells and entering other cells [62]. There is even evidence that the immune system may react to these virions [62]. Additionally, some proviruses have become part of normal biological functions. A striking example of this is the protein Syncytin-1, coded for by HERV-W [17]. Syncytin-1 is a cell-cell fusion protein, notably involved in placental development [17].

An evolutionary purpose for the integration of retroviruses may have been to offer protection from extant exogenous retroviruses. A current example of that may be seen in a mouse population at Lake Casitas in California, which suffers from an infectious degenerative neurological disease, a variant of murine leukemia virus (MLV). Some of the mice, however, are resistant against the disease. Those mice have an apathogenic endogenous MLV that is thought to bind and block the same cell surface receptor that the exogenous, pathogenic variant would use [15]. It is likely that such an effect would have occurred in humans as well at some point in time, and it is conceivable that it could even now be the case for some polymorphic HERV-K (HML-2) variant.

A review by Jonas Blomberg et al. lists other potential beneficial roles of HERVs, such as tissue specific enhancers[51], a polyadenylation signal and site [8], structural and nucleic acid binding proteins capable of packaging RNA, an envelope protein

which binds to a host surface protein, a spring-loaded transmembrane protein ready to fuse membranes and possibly cause immunosuppression. Blomberg even claims that a late part of the HERV life cycle may be to become "physiological servants" to their host as their negative effects are silenced, while potential beneficial roles ensure their survival in the genome. [10]

2.1.4 Associations with diseases

One of the main motivations of HERV research is to determine what direct or indirect pathological effects endogenous retroviruses could have. If such connections can be discovered and understood, it may yet help in the treatment of those pathologies. As formerly fully functional retroviruses, it would not be surprising for HERV proviruses to have retained some pathological potential. Direct detrimental effects of endogenous retroviruses have been documented in non-human animals. Species such as mice and sheep exhibit very recent ERV integrations, which still behave in the manner of the exogenous retrovirus from which they originate. An example of this is the intracisternal type A particle sequences (IAPs), an ERV that causes genomic damage in mice through a high transposition rate [40].

Though no human ERV has been shown to have such a direct cause in disease [57], many associations has been established. These include neurological diseases such as multiple sclerosis (MS) and schizophrenia, some cancers, and many autoimmune diseases such as diabetes and systemic lupus erythematosus [10]. For example, the aforementioned protein Syncytin-1, is also believed to be an immunogen involved in MS [17]. Notably, there are changes in HERV expression in cancer, and they are significant enough to possibly act as an early biomarker of some cancers [59]. High mRNA levels of various HERV Env genes (HERK-K, HERV-R, HERV-H) have been found in primary breast cancer patients compared to normal controls, and these levels were found to decrease with adjuvant chemotherapy [16]. A challenge in investigating such associations is their directionality. It is possible that HERV expression increases as a result of immune activity, while not being itself a cause of the activity [57].

Despite these associations with diseases, retroviruses with purely negative effects that cannot be silenced would have been unlikely to become fixed in the gene pool, as selective pressure would have worked against their endogenization. These pathological effects of proviruses are thus only likely to happen under specific conditions, where normal mechanisms of their silencing fail, such as in some diseases or as a result of age-related biological changes.

An additional factor to consider is the polymorphism of HERV-K (HML-2) in human populations [30]. The direct insertion of a provirus alone can cause disruption

of gene function in the integration locus. Therefore, this variance in provirus integrations is another potential source of pathology and could cause a difference between those who have the provirus and those who do not.

2.2 Proviral cargo of extracellular vesicles

Both beneficial and pathological roles of HERVs could be connected to extracellular vesicles (EVs). Retroviruses are thought to be closely associated with exosome formation and a viral infection can even alter the contents of exosomes, subverting their original function [4]. This link between retroviruses and extracellular vesicles could remain in endogenized retroviruses.

There is some evidence of such a connection, in disease states where proviruses are more freely expressed. EVs released by tumor cells have been found to contain cargo that is highly enriched in retrotransposon elements, including HERV mRNA [63]. This could be a method for proviruses to exert their influence when normal mechanisms of keeping them under control are not working. Investigation of EV cargo could therefore advance understanding of the potential biological significance of proviruses.

Extracellular vesicles as a whole have received renewed interest in recent years [42]. These vesicles, containing lipids, proteins, miRNA, and mRNA, are released by cells into the extracellular space. The vesicles can be transported to cells and their contents may even be translated into functional proteins in the case of mRNA, and affect gene expression in the case of miRNA. The purpose of this may be cell-to-cell communication. Much work is still needed to understand the biology of EVs. [55]

The RNA cargo of secreted exosomes largely reflect that of their parental cells [29]. Extracellular vesicles in circulation could originate from a variety of cell types [13]. It is thought that EVs could possibly act as biomarkers for a variety of age-related illnesses such as cancer, and neurodegenerative, metabolic and cardiovascular diseases [13].

2.3 Age-related changes in proviral expression

Various biological changes are associated with aging, including changes in the expression of HERV proviruses [32]. The potential of endogenous retroviruses to act as biomarkers of aging has been researched for example by Wada et al. already in 1993 [58]. They used PCR (Polymerase Chain Reaction) to study age-related ERV expression in mice and concluded that ERV expression in mice is "age-dependent" and

shows potential to be a biomarker of aging in humans as well [58].

If HERVs play a role in age-related decline, understanding that role could help in combating age-related disease. Aging is associated with a general decline in immune system function, referred to as immunosenescence [31]. Connected to this is the phenomenon of inflammaging; a low-grade, chronic, and systemic inflammation associated with aging [31]. Another notable age-related change is the so-called "senescence associated secretory phenotype", where senescent cells are no longer dividing, yet are still metabolically active and influencing surrounding tissue [55]. Furthermore, concentration of EVs has been found to decrease with aging [13]. With all age related changes, it can be difficult to distinguish whether they are causes or results of aging.

Some of these age-related biological changes are strong and distinct enough that there exists the phenomenon of the epigenetic clock: through measuring of DNA methylation the age of the subject can be determined to surprising accuracy [24]. The rate of this epigenetic aging is not constant, neither across lifespan nor among different individuals, and can be affected by factors such as the amount of body mass [35].

HERVs are also subject to this epigenetic change. HERV-K methylation levels have been found to be negatively associated with age [23] and upregulation of retroelements has been reported in aging and senescent cells [11]. It seems that as we age, the control mechanisms that keep our bodies working as intended, loosen, and as a result, these viral elements may become freed, with pathological repercussions.

2.4 The promise of bioinformatics methods

What unites all the biological phenomena described in this chapter thus far is that only little is known about them. The repetitive and mutated genomes of proviruses make them difficult to locate and to determine their potential biological significance. The small size and varied types of extracellular vesicles causes their analysis to be challenging. The undoubtedly complex interplay of these elements with aging remains to be unraveled. Past methods have proven inadequate in overcoming these challenges.

However, recent advances in genomics and bioinformatics have already enabled great developments in our knowledge of HERV biology [21]. Bioinformatics is a relatively new field of research in which a combination of computational, statistical and mathematical methods are used to study biology.

Quantitative real-time PCR has been the predominant method to study HERV expression [5], however RNA sequencing (RNA-seq) is now showing great potential

for HERV research [46]. RNA-seq enables quantification of HERV expression on the individual provirus level. RNA-seq methods are able to measure the transcription of regions of interest in the genome by quantifying the amount of mRNA originating from each specific region. The usual target is genes, yet RNA-seq can also be utilized in other tasks such as to study the expression of viral genes [64] or to measure the transcription of endogenous proviruses [19]. Alignment and annotation of RNA-seq data can be done with a splice-aware alignment tool. To be splice-aware is to be able to account for the introns that are present in the reference genome, yet missing from the RNA reads produced by RNA-seq.

Processing of high-throughput sequencing data is context specific, meaning that there is no single right way to go about alignment and annotation. Generally the goal is to achieve the maximum amount of high quality reads correctly aligned against the genome. However, what constitutes a "high quality read" is not always clearly defined and correct alignment can be ambiguous. A compromise between sensitivity and specificity is always required. Sensitivity, in any analysis, is the goal of maximizing the number of found true positives. Specificity, in contrast, is the goal of minimizing the number of false positives. The right balance of sensitivity and specificity is determined by the context of the analysis.

Proviruses can be studied with RNA-seq similarly to genes, though there are some difficulties. As repetitive elements, proviruses are less differentiated than human genes. This means that mapping transcripts originating from proviruses to correct genomic loci can be challenging [19]. Another consideration is that HERV expression is thought less likely to lead to functional protein products than normal gene expression, as many HERVs are thought too mutationally decayed for that. Thus, HERV transcription may not translate to HERV expression as directly as one might otherwise expect. Yet, proviruses can indirectly affect the expression of neighboring genes, for example through their LTRs [26], and thus their amount of transcription is still valuable information.

Though it is very early to discuss any clinical applications, when we know so little of HERVs, the potential for such applications can already be seen. Though it is still unclear if HERVs can drive diseases in humans, the expression of HERVs has been shown to change in different contexts. With a better understanding of the associations between HERVs and disease, proviral expression could perhaps be used to diagnose diseases that might be difficult to identify through other means. For example, change in HERV expression patterns may be an early sign of cell transformation and thus could be used as a biomarker for cancer [19].

If proviral expression does indeed change with age, it could also work as an aging

biomarker. Consequently, it could be used to determine an individual's biological age, which is a more accurate predictor of age-related issues than is chronological age [24]. Risk for development of age-related diseases could be identified earlier, so that treatment can be begun sooner and have a greater effect. Accurate measures of biological age could have many applications, some of them outside health care, such as to verify the age of someone with no documentation of it.

Identifying disease states in which HERVs are involved in and understanding how HERVs are involved, could alone increase understanding of those disease states. HERV research could similarly lead to a better understanding age-related decline and how to combat it.

There is now a need for smaller exploratory studies to investigate different approaches and to provide information on which seem most promising. Those areas of research may then be focused on in larger, more resource-intensive studies. In the chapter that follows, I will outline the concrete deliverables that this exploratory thesis work could contribute to this field.

3. Objectives

The focus of this study was to investigate proviral expression in EVs from plasma samples, contrasting that to proviral expression in PBMC samples derived from the same individuals. Simultaneously, the proviral expression has been compared between two distinct age groups of the study participants. Therefore, the data can be divided into four groups based on the two biological sources of PBMCs and plasma as well as age groups of nonagenarians and young controls, as demonstrated in table 3.1. More information on sample selection and used datasets is provided in the next chapter.

Table 3.1: The four sample categories that are characterized and compared in this work.

PBMC nonagenarian	plasma nonagenarian
PBMC young control	plasma young control

The aim of this exploratory research work has been to characterize and compare the four HERV expression profiles formed by this categorization, in order to guide the direction of later studies. The driving research questions were along similar lines: What are the characteristics of the PBMC and plasma RNA HERV profiles? How do these profiles differ between each other and across nonagenarians and young controls?

These comparisons were not the sole target of this work. The question of proviral expression itself, whether it exists and how much, is still in contention. Do our results support the hypothesis that there is proviral expression in PBMCs and plasma? Another concern is technical. Does our RNA-seq data analysis pipeline appear to work? Can HERVs be investigated in this way? We have also been on the lookout for proviruses of interest that could be studied further.

4. Materials

The study is based on data on the expression of genes and HERV proviruses quantified through RNA-seq of PBMC and plasma samples. This chapter describes all the steps that were taken to produce those datasets, from initial sample collection to final preprocessing and normalization of the resulting read counts. Filtering of plasma samples and the RNA-sequencing of all samples was done by Institute for Molecular Medicine Finland (FIMM). Quality control, alignment, expression quantification, and normalization was done by Genevia Technologies.

4.1 Sample collection

Blood samples were drawn from young and nonagenarian participants. There were seven young controls aged between 26 and 32, with a median age of 28. The young controls were healthy laboratory personnel, non-smokers with no chronic illnesses and no infections or vaccinations within the two weeks prior to blood sample collection. The case samples were provided by seven nonagenarians, participants in the Vitality 90+ Study, who were carefully selected to be representative of the median of both cognitional and functional fitness. The Mini Mental State Examination (MMSE) was used to determine mental capacity of nonagenarians [1], with a median score of 23 for the full cohort and 22 for the chosen samples. Barthel scale [44] was used to ascertain physical capability of nonagenarians, with a median score of 85 for both cohort and chosen samples. All participants are female, as aging-associated changes in gene expression have been found to differ between males and females [31]. Including samples from males could have detrimentally affected the statistical analyses.

The low number of replicates may limit the conclusions that can be drawn from the analyses done on the data, yet the purpose of this work is to explore the potential of the chosen approach. For such purposes the number of replicates was deemed high enough. [43]

4.2 RNA extraction

Whole blood samples from nonagenarians and young controls were used to extract PBMC and plasma for RNA-seq. Purification of the RNA used for RNA-seq was done with a miRNeasy mini kit (Qiagen, CA, USA). Quality of the RNA, in terms of concentration and quality was validated with a NanoDrop ND-1000 spectrophotometer

(NanoDrop Technologies, Wilmington, DE, USA).

The plasma samples were found to be of high quality still despite a long storage time. The structure of exosomes in plasma protects the genetic materials contained within, as high quality mRNA has been reportedly extracted from sera stored for over a decade [39]. The presence of intact mRNA in the samples is thereby in itself evidence that the quantified expression originates from extracellular vesicles, since extracellular mRNA free in the blood would have been much less likely to survive the storage.

Use of PBMCs to study proviral expression does introduce one confounding factor that is sample cell type heterogeneity. Peripheral blood mononuclear cells include a variety of cell types, which could exhibit differing proviral expression. Majority of the cell population of PBMCs are lymphocytes, such as natural killer cells, T cells, and B cells, though some of the sample mRNA can also be expected to be from monocytes.

4.3 Filtration of plasma samples

Filters of size 0.8 µm were used to remove cell debris, platelets and other larger particles from the plasma samples in order for them to portray more accurately the extracellular vesicle mRNA content. Filtering was done by Institute for Molecular Medicine Finland (FIMM). The filtering was reported to leave only smaller RNA containing units, such as exosomes and other extracellular vesicles, viruses, and possibly some lipoproteins. Thus the main targets of study, exosomes and other extracellular vesicles, are represented to a greater degree in the RNA sequencing. Possible problems after filtering are a lack of rRNA making QC more challenging or just a general lack of RNA, yet this was reported by FIMM not to be the case with these samples. Both unfiltered and filtered plasma were sequenced to verify that the filtration worked as expected.

4.4 RNA sequencing

The RNA sequencing was done by the Institute for Molecular Medicine Finland (FIMM). Integrity of total RNA was evaluated with Agilent Bioanalyzer RNA nano chips (Agilent). RNA in the samples was quantitated with Qubit RNA–kit (Life Technologies). Ribodepletion of rRNA was done with 1 ug of total RNA utilizing ScriptSeq™ Complete Gold System (Epicentre) as well as for RNA-seq library preparation. SPRI beads (Agencourt AMPure XP, Beckman Coulter) were used for purification of RNA-seq libraries. Evaluation of the library QC was done on High Sensitivity chips by

Agilent Bioanalyzer (Agilent). Paired-end sequencing of RNA-seq libraries was done using Illumina HiSeq technology with a minimum of 60 million 2x100bp paired-end reads per sample. The resulting raw RNA-seq data in the format of fastq-files was forwarded to Genevia Technologies for further processing.

4.5 Quality control and read trimming

The quality of the sequencing output was evaluated with the tool FastQC. Not all samples are necessarily of high quality. If there are any significant technical problems with a sample, it should be filtered out. Even high quality samples usually contain reads of varying quality. Low quality reads may be trimmed or filtered out.

Though the RNA sequencing data was mostly of good quality, one nonagenarian and one young control sample were discarded as advised by Genevia. Some trimming and filtering of the remaining samples was needed. FastQC results guided the trimming of reads that was done with the tool "Trim Galore!". A quality score of 20 was set as the general cutoff. Trimming of both ends of reads was tailored individually for each sample based on FastQC results. After filtering and quality control the data was ready for alignment and annotation.

4.6 Alignment and annotation

The tool TopHat v2.0.13 [52] was used for alignment of the reads against a HERV and gene transcriptome reference, with default parameters. It should be noted that TopHat is deprecated and STAR or other newer aligners might perform better. TopHat requires as input RNA-seq data as FASTQ or FASTA, a reference genome and an annotation file. Annotation of aligned reads to proviral sources is done similarly as for genes. Genomic loci of proviruses is all the information that is needed. HERV-K (HML-2) annotation data was obtained from Subramanian et al. [47] and HERV-W from Grandi et al [17].

Human genome reference build 19 (hg19) was used as the reference in alignment. The reference genome is not derived from any single individual, but is rather an amalgamation of data from many individuals, modeling the shared genome. As HERV-K (HML-2) proviruses are thought to be polymorphic in the human population [30], some may be absent from the reference genome, yet present in the genomes of study participants.

Reads mapping to multiple regions of the genome (non-unique) were filtered out with SAMtools. One potential issue with this approach is that quantifying only

unique reads biases the results against newer, less diverse HERV integrations [19]. This is due to older integrations having become more differentiated and therefore producing a greater number of unique reads [19]. This research focuses on HERV-K (HML-2), which is the most recent integration and thus the least differentiated. This may cause a great deal of otherwise good reads to be thrown out. However, the expression of different HERV families is not compared, which would cause non-unique read bias to affect the results. Ultimately, accurate mapping was prioritized over this potential issue.

Cufflinks2 v. 2.2.1 [54, 53] was used for two data processing steps. Firstly, to ascertain raw expression estimates with Cuffquant and then to perform geometric normalization with Cuffnorm.

4.7 Normalization

The result of alignment and annotation is read counts for each location of interest, in this case genes and proviruses. Lengths of genes and proviruses vary, however, meaning that a longer transcript is more likely to have a higher number of counts originating from it. For these counts to be accurate within a sample, the counts should be adjusted to account for the length of each transcript. Additionally, in order for these read counts to be comparable across samples, they should be normalized together.

The tool Cuffnorm from the Cufflinks toolset [54, 53] was used for these steps. Cuffnorm geometric normalization was utilized, which performs normalization according to the method originally outlined by Anders and Huber [3]. FPKMs and read counts are scaled for each sample by the median of the geometric means of read counts across all samples. For purposes of better representative normalization, expressions of HERV elements were quantified and normalized together with ENSEMBL v. 82 gene reference set.

4.8 Ethics

Written informed consent was provided by the study participants. The research was conducted according to the principles expressed in the declaration of Helsinki. The study protocol was approved by the ethics committee of the city of Tampere (1592/403/1996).

5. Methods

This chapter describes the analyzes that were done using the normalized read count data of genes and proviruses. The materials section outlined procurement and processing of datasets done prior to this thesis work, while this chapter focuses on the methods used during the thesis work itself. Identified limitations of the methods of this study will also be clarified and discussed.

5.1 Investigation of proviral expression

Analysis of differential expression of proviruses between conditions, in this case age groups, can be done similarly as it is done for genes. The Mann-Whitney U test was used to determine the significance of differential expression.

It is difficult to distinguish between no expression and low expression of a gene or provirus, based on RNA-seq data. However, a cutoff of normalized read count of 16 has been used in previous work on this PBMC dataset to separate expressed from the not expressed proviruses. Thus, only proviruses that exhibit a normalized read count of at least 16 in the majority of both young control and nonagenarian samples were considered expressed when determinng significant differential expression. The use of this cutoff was continued in this work for the purposes of continuity and comparability.

The multiple testing problem is important to account for when testing for significant differences in the expression of a large number of genes and proviruses. When enough tests for significance are made, some of them will by chance result in significant p-values even if none are truly significant. Strictly speaking a p-value of 0.05 means that there is a 5% chance to arrive at this result if the null hypothesis is true. Therefore, if a thousand genes are tested for significantly differential expression, and 0.05 is chosen as the cutoff of significance, one would expect 50 of those genes to be reported as significantly differentially expressed, even if none are in reality. To control this false discovery rate, the Benjamini-Hochberg method [9] has been utilized in all analyses that result in p-values. In the Benjamini-Hochberg method, the number of expected false positives (false detection rate) is used to adjust p-values to ascertain significance even with a very large number of tests. [33]

It should be noted that statistically significant differences do not necessary translate to biologically significant differences. Fold change between age groups was used as an additional measure to predict biological significance. Fold change is the difference in amounts between two conditions. Log 2 fold change better visually represents

upregulation and downregulation than does raw fold change. This is due to the division of log 2 fold changes into positive and negative values, with higher absolute value indicating greater difference. Whereas, with raw fold change upregulation is shown by positive numbers, while downregulation is limited to between 0 and 1. In this work fold change is only ever used as a guide and never as a strict cutoff, as even a seemingly small difference could potentially have large biological significance.

Spearman correlation was predominantly used in this work to evaluate ranked correlation. The differences in PBMC and plasma sample contents could discount the linear correlation of Pearson coefficient, yet Spearman only evaluates the correlation of ranked expression of genes and proviruses. If the same genes and proviruses are ranked highly, and the same lowly, then there is correlation. This is irrespective of the differences in the magnitude of proviral expression, as is seen between PBMC and plasma samples.

5.1.1 Scaling of proviral expression with chosen housekeeping genes

As the expression seen in plasma samples was generally lower than in PBMC and varied greatly, scaling was done with housekeeping genes in order to correct for potential biases. Housekeeping genes are genes involved in basic and vital functions for cells and they are thus universally expressed. Even designated housekeeping genes can vary in expression across tissues. Selection of correct housekeeping tasks for each analysis is crucial as choosing wrong housekeeping genes can skew the results. [45]

A panel of commonly used housekeeping genes, consisting of GAPDH, ACTB, GUSB, and HPRT1 was used here, due to individual housekeeping genes having very varied expressions. The mean of the housekeeping panel was used to scale gene and proviral expressions.

5.2 Proviral expression patterns

The total proviral expression and differences in the expression of individual proviruses do not tell the whole story. There can also be differences in expression patterns between age groups that are visible only when studied as a whole. Such as correlations between genes and proviruses in how they are expressed in different samples. These could indicate relationships between proviruses and genes, for example related functions or biological pathways. Principal Component Analysis (PCA) as well as hierarchical clustering were used to study such potential patterns.

5.2.1 Principal component analysis

The high dimensionality of sequencing data can make it difficult to interpret. There are many genes and many samples. PCA is a data scaling method to compact high dimensional data into fewer dimensions to aid in the interpretation of data. PCA analysis was done in R utilizing the prcomp-function. PCA was performed separately for both HERV-K (HML-2) and HERV-W. PCA is a quick way to see if there are differences, yet is limited in what can be deduced from it. Hierarchical clustering was done as another exploratory analysis to expand and contrast the PCA.

5.2.2 Hierarchical clustering

Hierarchical clustering of the samples based on normalized provirus read counts was done to investigate if there are subgroups of samples with similarities in their expression profiles [20]. In such agglomerative hierarchical clustering, the most similar samples or clusters are joined into a new cluster at each step, with similar samples falling into the same cluster [12]. From the resulting hierarchical structure, one can then discern notable clusters contained within. The clustering of the expression profiles was done to ascertain whether they would cluster differently between age groups.

Many factors can affect the clustering results. Clustering is based on measures of similarity, and thus the chosen definition of similarity will determine how the samples cluster together. How the data was preprocessed can affect similarity measures and thus clustering [12]. What distribution the data points follow can be important, as certain similarity measures assume certain distributions [12].

The accuracy of clustering results will vary greatly depending on the chosen methods and it is therefore recommended to try multiple different clustering algorithms and parameters [12]. There is no single right approach to clustering [12]. Instead, the used methods should be chosen based on the type of data and the context of the clustering. Although one can evaluate the uncertainty involved in clustering [48], clustering is usually used as an early, exploratory step in analysis [12], and other methods can be used to further investigate and verify the results.

Hierarchical clustering of the samples based on normalized read counts was done separately for both HERV-K (HML-2) and HERV-W. Spearman correlation was used as the distance metric, which is robust against outliers and non-Gaussian distributions, and can capture nonlinear relationships [2, 27]. Ward's minimum variance was used as the linkage method, which has been reported to perform better with gene expression data than the more traditional methods of average and complete linkage [2]. Multistep-multiscale bootstrap resampling was done using the R package Pvclust to

evaluate the uncertainty involved in the clustering [48]. Thousands of samples of varying sizes are randomly created from the data and then clustered. An approximately unbiased (AU) p-value is obtained, which indicates the bias corrected percentage of dendrogram variants where the specific cluster was observed.

Clustering of PBMC samples was done with all the proviruses that had a normalized read count of the established cutoff of at least 16 in at least one sample. With plasma samples all proviruses that had at least some expression (normalized read count > 0) were used in clustering.

5.3 Inferring potential biological effects of proviruses

The aim of HERV research is not only to examine and characterize proviral expression. The goal is to determine to what extent proviruses can affect physiology and in what ways. Gene ontology enrichment analysis was used to investigate potential biological significance of HERV proviruses. In the case of genes, one method to infer possible functions is to explore the known functions of other genes with similar expression patterns [28]. These known functions can be retrieved from the Gene Ontology (GO) database. The GO project is an ongoing effort to collect information on the biological processes, molecular functions and cellular components that genes are involved in. In the context of GO, biological process specifically refers to a "biological objective to which the gene or gene product contributes" that is achieved through "ordered assemblies of molecular functions" [50]. The GO database is not limited to a listing of processes, instead containing a hierarchical structure of relationships between processes. [50, 49]

Knowledge of HERV proviruses is far too limited for there to be a comparable database of provirus ontology. Yet proviruses act analogous to genes in many respects and the same approach can be extended to investigating the biological processes they might affect. Proviral co-expression with genes that have known functions could indicate involvement in similar biological functions.

GSEA (Gene Set Enrichment Analysis) was used in this work to infer links between proviruses and GO terms. GSEA can be done based on a list of ranked genes. Here the genes were ranked according to correlation in expression to a provirus, to infer what functions the provirus in question could be involved in. A similar approach has been adopted for example by Zhang et al., when studying co-expression of Epstein-Barr virus genes with human genes to identify potentially affected human biological pathways [64].

One difficulty with this approach is that unlike genes, most proviruses are not

necessarily expected to have any function. Yet when comparing many proviruses with many genes, there will be incidental correlations in expression. This can lead to inflated p-values and a large number of false positives compared to few true positives.

5.4 Investigation of genomic neighborhood

In addition to searching through literature for more information on proviruses of interest, different extended methods were also utilized for more in-depth study. Integrative Genomics Viewer [41] was used to study the genomic landscape surrounding such proviruses. The location of a provirus is of interest, because it could be located next to, or even within, a gene. If the gene is highly expressed, this could explain a high expression of that provirus. Furthermore, a provirus could affect the expression of any nearby genes. Online databases and tools, such as GTEx (Genotype-Tissue Expression) [18], were then used to follow up on locational information to investigate if anything of interest can be found.

6. Results and discussion

After determining that there indeed is proviral expression in both PBMC and plasma samples, a variety of analyses were performed to characterize and compare this expression between the biological sources of PBMCs and plasma, as well as between the age groups of young controls and nonagenarians. In this chapter, the results of the thesis work are presented and discussed analysis by analysis.

Before proceeding with more specific analyzes the overall format of the data was investigated. Final format of the given materials is as matrices of normalized read counts. There is one dataset for PBMC samples and another for plasma samples. Both contain data from the young controls and the nonagenarians. Rows of the matrices are genes and proviruses, while columns are for different samples.

Of note is that only an extremely small portion of total quantified mRNA originates from proviruses. An overwhelming majority of it is derived from human genes, as shown in table 6.1. Proviral expression constitutes a slightly larger portion of all expression in PBMC than in plasma, though the difference is negligible. There was significant differential expression of many human genes between age groups in PBMC data, as determined by Benjamini-Hochberg corrected p-values lower than 0.05. The focus of this work is proviruses, yet this data could also be used to study age-associated differences in gene expression. There was less differential expression between age groups in plasma, where there were only 19 genes and no proviruses differentially expressed.

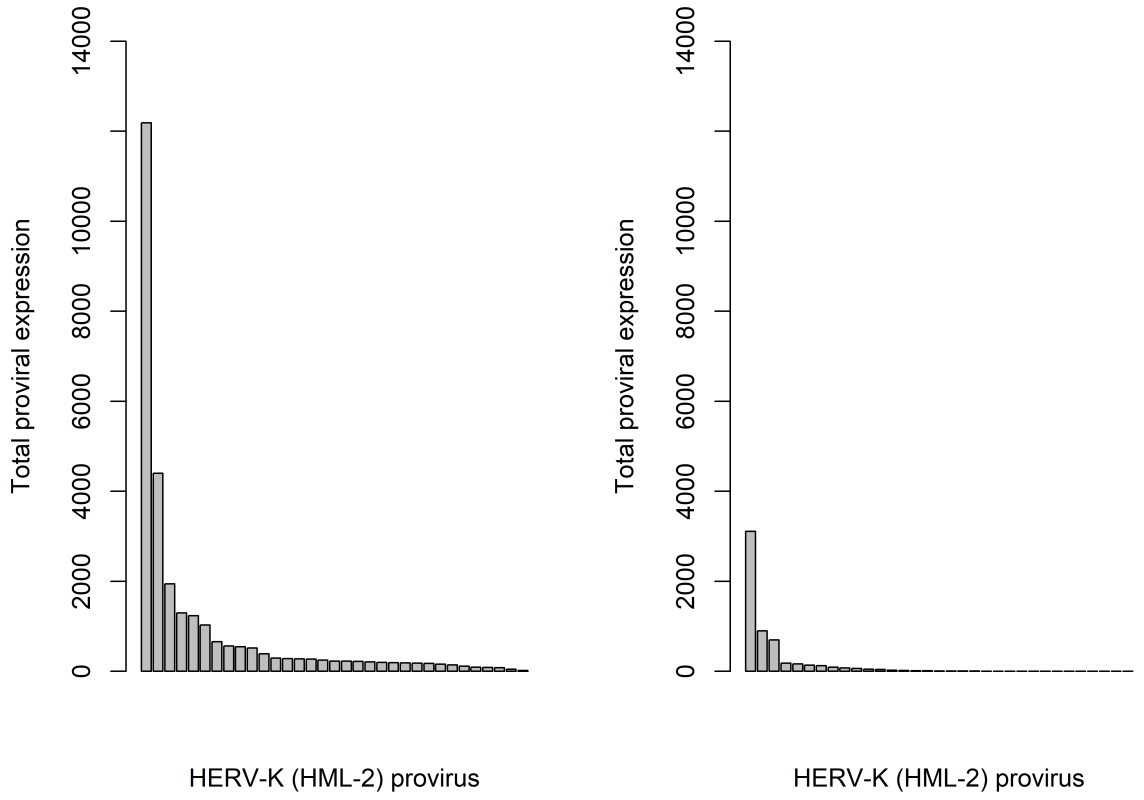
Table 6.1: Percentage of expression from total quantified expression.

Genomic entity	PBMC	Plasma
Human genes	99.968 %	99.980 %
HERV-K (HML-2) proviruses	0.011 %	0.007 %
HERV-W proviruses	0.021 %	0.014 %

6.1 Proviral expression in blood cells compared to plasma

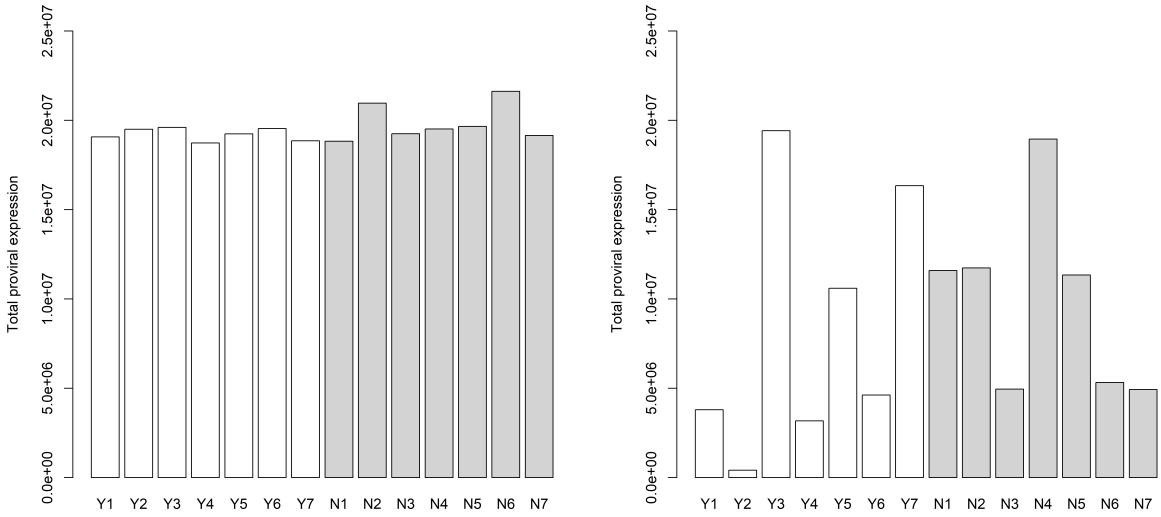
Both the overall scale of expression, as well as what proviruses are expressed, were found to be very different between samples derived from PBMCs and from plasma. Figure 6.1 illustrates how PBMCs had much higher overall normalized read counts than plasma. The total normalized read count additionally greatly varied from sample to sample in plasma samples, yet not in PBMC samples, as shown in figure 6.2. These

differences were taken into account in following analyses by focusing on the relative expression of proviruses, instead of the absolute read count.



(a) PBMC sample read counts by provirus. **(b)** Plasma sample read counts by provirus.

Figure 6.1: The problem of scale. Bar plots illustrating how normalized HERV-K (HML-2) read counts from PBMCs and from plasma show strikingly different overall expression levels. Shown are combined read counts per provirus from both young and nonagenarian samples, sorted in order of highest expression separately for PBMC and plasma.



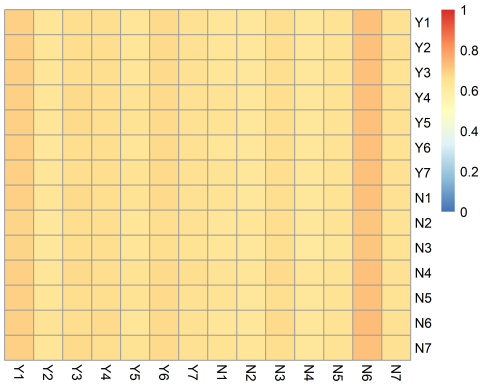
(a) PBMC sample read counts by sample. **(b)** Plasma sample read counts by sample.

Figure 6.2: Bar plots showing the difference in total number of HERV-K (HML-2) normalized reads between PBMC and plasma per sample. Samples from young controls are denoted by the letter Y, and nonagenarians by the letter N.

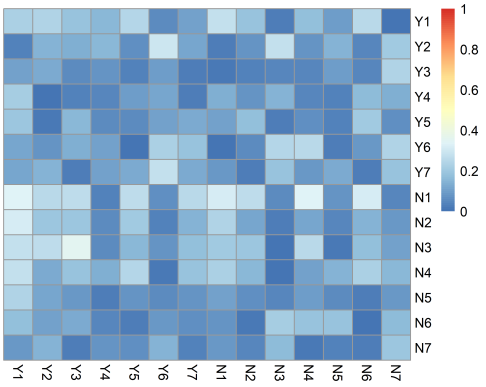
Scaling of expression values based on a panel of housekeeping genes was tried, yet the results appeared arbitrary and unreliable. Instead, methods that do not rely on comparable absolute expression were used in this work, such as Spearman correlation coefficient.

The expression of individual genes and proviruses was compared between PBMC and plasma samples and it was found that the expression of genes correlated more strongly than did proviral expression. Spearman correlation coefficient for genes between PBMC and plasma was 0.64, indicating moderate to strong correlation. For HERV-W proviruses, the coefficient was 0.23 indicating weak to no correlation. For HERV-K (HML-2) proviruses the coefficient was -0.0025, indicating virtually no correlation.

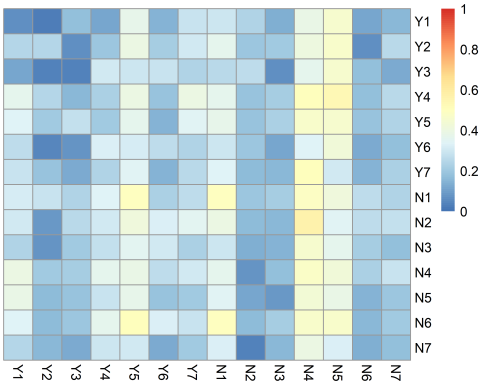
Sample to sample correlation between PBMC and plasma expressions is illustrated in figure 6.3. Nonagenarian participant N6 had higher correlation in gene expression between PBMC and plasma than the others. Nonagenarian participants N4 and N5 had higher correlation in HERV-W proviral expression between PBMC and plasma. Whereas HERV-K (HML-2) expression was very uniformly of low correlation between PBMC and plasma.



(a) Human gene expression



(b) HERV-K (HML-2) proviral expression



(c) HERV-W proviral expression

Figure 6.3: Heatmaps of sample to sample correlation of expression values between PBMC and plasma using Spearman correlation coefficient. Produced using R package pheatmap.

Fold changes between PBMC and plasma for each provirus were also calculated and are shown in tables 6.2 and 6.3. Direct comparison of within experiment normalized read counts between experiments is problematic, and thus these fold changes are only shown to convey overall, relative differences between PBMC and plasma samples, and not as a absolute measurement of difference in expression for each provirus in the two conditions.

The plasma samples have been filtered to consist mostly of extracellular vesicle contents, and enrichment of proviral cargo in EVs has been reported, though in cancer [63]. Therefore, this discrepancy in PBMC and plasma expression between proviruses does support the hypothesis of proviral contents being selected for EV cargo. Of note is that HERV-K (HML-2) shows greater difference between biological sources. Following analyses will show that HERV-K (HML-2) was also found to be expressed differently between age groups. If selective loading of proviruses for EV cargo does occur, it might not be universal, but rather focused on certain families, subfamilies or individual proviruses. It is nevertheless important to remember that the expression of HERVs could be different for many reasons, and that these results alone cannot

Table 6.2: HERV-K (HML-2) median expression in young controls, as measured by normalized read count, in PBMC and plasma as well as the fold changes calculated from them. The table is ordered first by fold change and then by median expression in PBMC. No fold change was calculated for proviruses, which had zero expression in either PBMC or plasma.

HERV-K locus	Median expression in PBMC	Median expression in plasma	Fold change
19q13.12a	18.10	133.54	7.38
19q11	4.91	9.42	1.92
10q24.2	8.95	6.28	-1.42
20q11.22	8.57	4.32	-1.99
9q34.3	3.81	1.40	-2.72
3q21.2	19.72	1.40	-14.07
8p23.1c	22.86	1.26	-18.18
11q12.3	8.93	0.45	-19.90
4p16.3a	16.20	0.45	-36.08
1q22	261.01	5.66	-46.13
19q13.12b	146.49	2.80	-52.25
9q34.11	36.90	0.29	-128.19
3q12.3	916.86	1.40	-654.09
12q24.33	97.25	0.00	
1q23.3	78.79	0.00	
7q34	74.16	0.00	
1q32.2	42.30	0.00	
14q11.2	27.67	0.00	
4p16.1a	26.10	0.00	
1q21.3	18.72	0.00	
10p14	18.00	0.00	
8p23.1b	17.09	0.00	
11q12.1	14.65	0.00	
11p15.4	12.35	0.00	
3q13.2	11.47	0.00	
8p23.1a	11.39	0.00	
19q13.41	9.71	0.00	
4p16.1b	9.44	0.00	
1p31.1a	6.30	0.00	
12q24.11	4.88	0.00	
22q11.21	4.76	0.00	
16p13.3	4.41	0.00	
12p11.1	0.00	0.47	

Table 6.3: HERV-W median expression in young controls, as measured by normalized read count, in PBMC and plasma as well as the fold changes calculated from them. The table is ordered first by fold change and then by median expression in PBMC. No fold change was calculated for proviruses, which had zero expression in either PBMC or plasma.

HERV-W locus	Median expression in PBMC	Median expression in plasma	Fold change
18p11.31	9.63	9.81	1.02
12q24.31	37.27	18.22	-2.05
11q14.2	4.69	1.40	-3.34
10q24.1	25.16	6.16	-4.08
4q21.22	15.45	3.45	-4.47
3q13.31	182.88	38.98	-4.69
3q26.32	12.71	1.89	-6.74
1p22.2a	26.36	2.06	-12.78
2q11.2	27.27	2.05	-13.27
3q13.32	58.60	3.59	-16.31
6p22.3	30.80	1.47	-20.99
1q22	11.82	0.46	-25.70
2q31.2a	31.76	0.92	-34.54
6q27b	36.58	1.03	-35.61
2q32.3	13.64	0.29	-47.37
3q23b	73.15	1.24	-59.09
1p34.2	48.72	0.47	-103.33
2p16.2	89.99	0.46	-195.73
2q22.2	113.64	0.00	
6q21a	111.17	0.00	
6q21c	38.13	0.00	
11q14.1	36.36	0.00	
Xp11.21	34.60	0.00	
1q42.13	33.35	0.00	
14q21.2	31.63	0.00	
17q22	25.42	0.00	
7q21.2	17.77	0.00	
19q13.2a	17.05	0.00	
9p13.3	16.31	0.00	
2p23.1a	16.17	0.00	
2q24.3	14.38	0.00	
13q13.3	13.64	0.00	
1q32.1	11.82	0.00	
4p16.3	11.72	0.00	
1p12	10.59	0.00	
6q24.2a	10.00	0.00	
17q12a	10.00	0.00	
8q21.11	9.18	0.00	
17q12b	9.09	0.00	
15q21.3	8.89	0.00	
14q32.11	7.06	0.00	
3q11.2	6.67	0.00	
5q22.2	4.49	0.00	
7q31.1a	1.93	0.00	
7p14.2	0.48	0.00	

determine the cause.

6.2 Differences in expression between age groups

There was an unexpectedly small difference in the total expression amount between age groups in both PBMC and plasma. An expected general upregulation of proviruses in samples derived from elderly participants was not observed. Distribution of read origin separated by age group is shown in table 6.4.

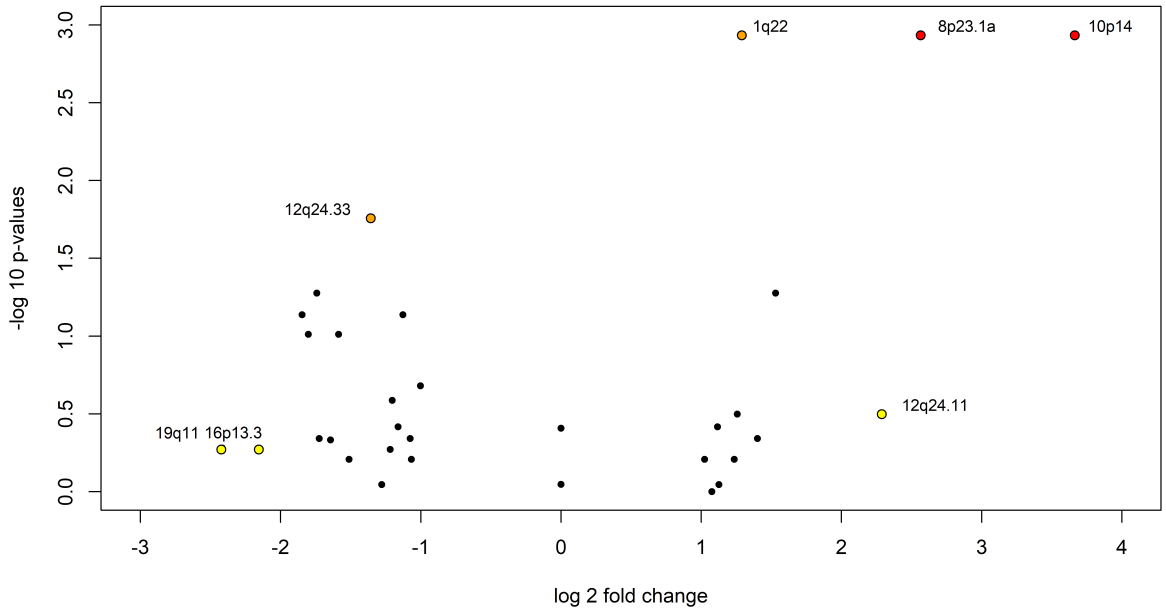
Table 6.4: Percentage of expression for each category of genomic entity from total quantified expression, separately for both age groups.

Genomic entity	PBMC		Plasma	
	Young	**Old**	**Young**	**Old**
Human genes	99.968 %	99.968 %	99.983 %	99.977 %
HERV-K (HML-2) proviruses	0.012 %	0.011 %	0.005 %	0.008 %
HERV-W proviruses	0.020 %	0.021 %	0.012 %	0.015 %

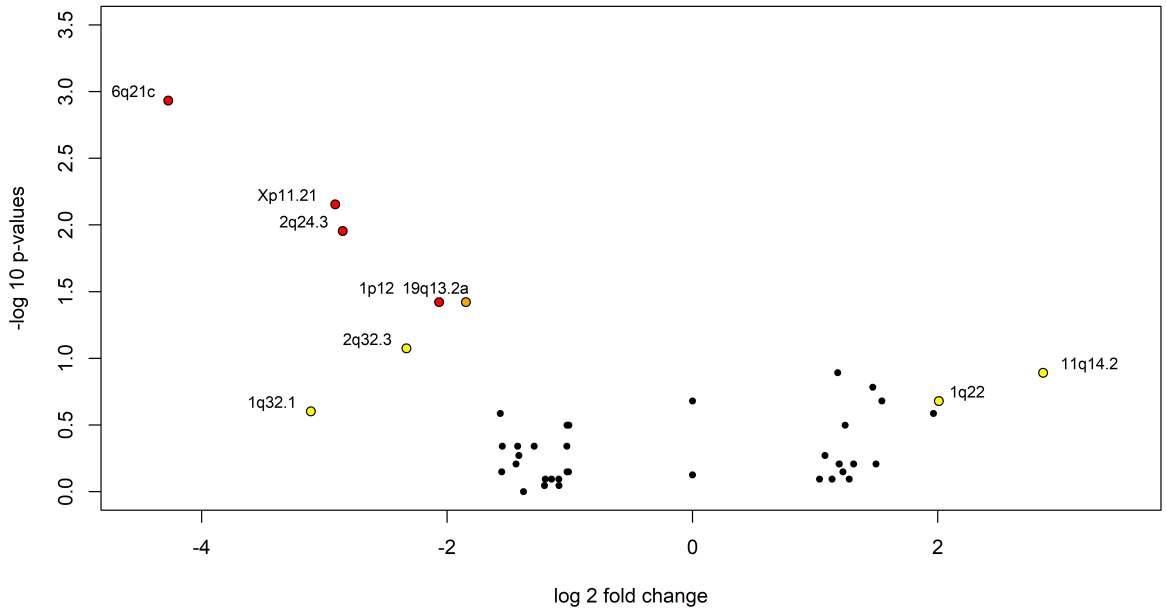
Some proviruses were significantly differentially expressed across age groups in PBMC samples, but none in plasma samples, as determined by Benjamini-Hochberg corrected p-values lesser than 0.05. Many proviruses had high median fold changes between age groups. Most HERV-K (HML-2) proviruses that were differentially expressed between age groups showed higher expression in the elderly participants. In comparison, differentially expressed HERV-W proviruses unexpectedly showed downregulation in the elderly.

Volcano plots that plot significance against fold change are shown in figure 6.4. Some of the proviruses on the plot that showed promise with unadjusted p-values were discounted by Benjamini-Hochberg p-value correction or by having a low overall normalized read count (previously established cutoff of at least 16 in majority of both young control and nonagenarian samples).

After filtering out proviruses with low expression and performing Benjamini-Hochberg p-value adjustment for multiple testing, four proviruses were found to be significantly differentially expressed in PBMC samples (p-value < 0.05), displayed in figure 6.5. Three of those were HERV-K (HML-2) proviruses 1q22, 10p14 and 12q24.33 and one HERV-W provirus Xp11.21. Provirus 1q22 is known by aliases K102, K(C1b), K50a, and ERVK-7 [47]. Provirus 10p14 is also known as K(C11a), K33, and ERVK-16 [47]. No aliases were found for provirus 12q24.33 or Xp11.21, besides the genomic loci used to distinguish them in this work.



(a) HERV-K (HML-2) proviruses



(b) HERV-W proviruses

Figure 6.4: Volcano plots showing significance plotted against fold change, when comparing age groups based on PBMC samples. The p-values shown here have been converted to -log 10, and fold changes are as log 2. Thus higher y-axis position indicates higher significance, while greater distance from 0 on the x-axis indicates greater change in expression values between conditions. A positive fold change here signifies upregulation in nonagenarians, while a negative value signifies downregulation. Dots colored yellow indicate proviruses with high fold change (absolute log 2 fold change higher than 2), orange indicate high significance (unadjusted p-value lower than 0.05), while red dots indicate both high significance and fold change.

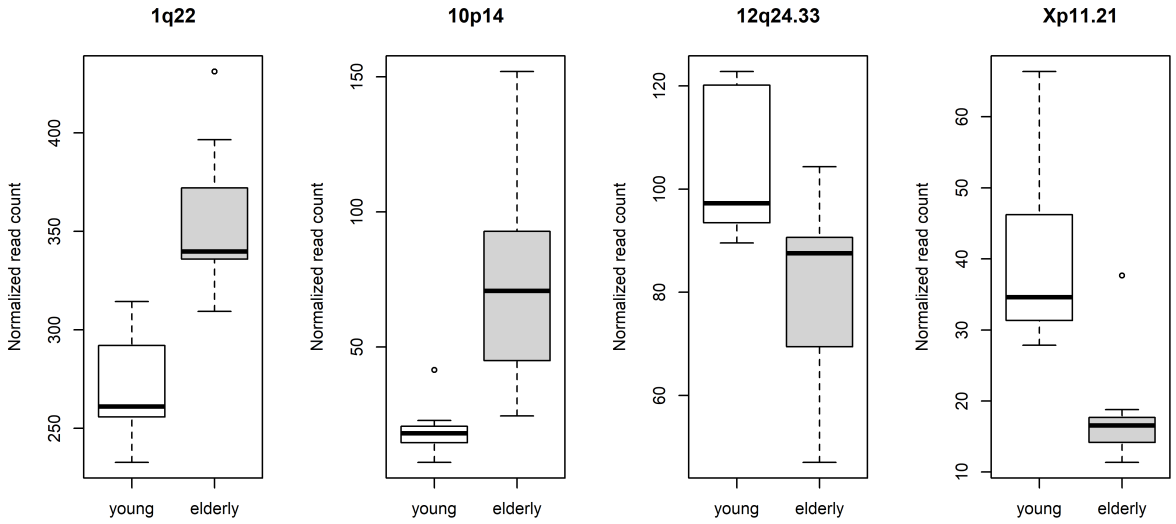


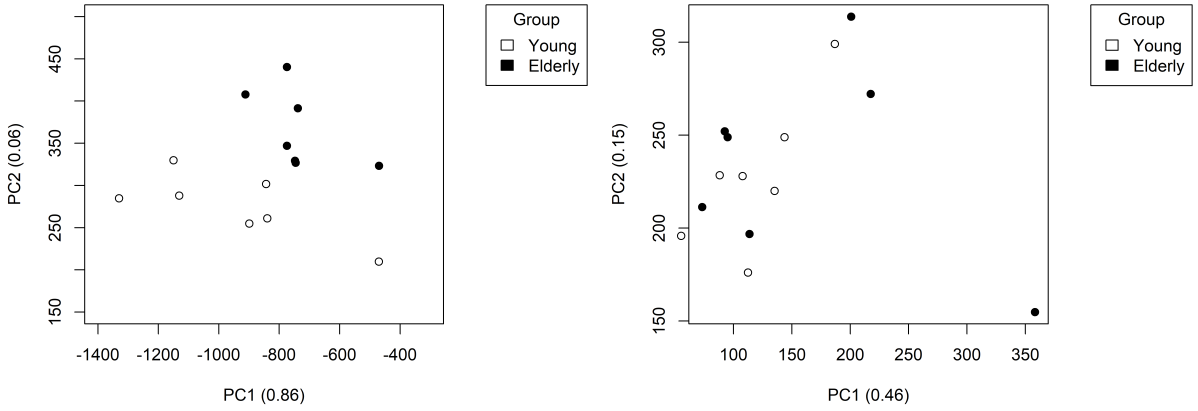
Figure 6.5: Significantly differentially expressed proviruses between young controls and nonagenarians. Proviruses 1q22, 10p14 and 12q24.33 belong to the HERV-K (HML-2) family, while Xp11.21 belongs to the HERV-W family.

Differences in proviral expression between age groups were lesser than anticipated, especially with plasma samples. It is possible that the low number of replicates is the cause and that with a higher number of samples the age group differences would be more prominent and less overshadowed by individual differences.

6.2.1 Patterns of expression across proviruses

Results from the Principal Component Analysis suggest that only HERV-K (HML-2) in PBMC samples exhibits age-associated expression patterns. Plot of the first two principal components from PCA analysis of HERV-K (HML-2) expression in PBMC samples showed two clusters, shown in figure 6.6. PCA of HERV-W expression in PBMC and of both provirus families in plasma samples showed no clustering at all.

Hierarchical clustering of PBMC samples produced results that mostly align with PCA. Clustering indicated clear differences in the expression profiles between nonagenarian and control samples with HERV-K and to a lesser extent with HERV-W, as displayed in figure 6.7. Clustering of samples based on HERV-K (HML-2) expression resulted in two groups separated along the age group lines. The clusters have approximately unbiased (AU) p-values of 98 and 97, which are equivalent to p-values of 0.03 and 0.02 respectively, indicating statistical significance (< 0.05). An AU p-value is the bias corrected percentage of resampling dendrogram variants where the specific cluster was observed. Hierarchical clustering of HERV-W did divide the samples into the two age groups, though the estimated significance of those clusters was lesser



(a) PBMC HERV-K (HML-2)

(b) PBMC HERV-W

(c) Plasma HERV-K (HML-2)

(d) Plasma HERV-W

Figure 6.6: Principal component analysis of HERV-K (HML-2) and HERV-W proviral expressions in PBMCs and in plasma. The proportion of variance for each plot that the two principal components account for are shown on the axis labels.

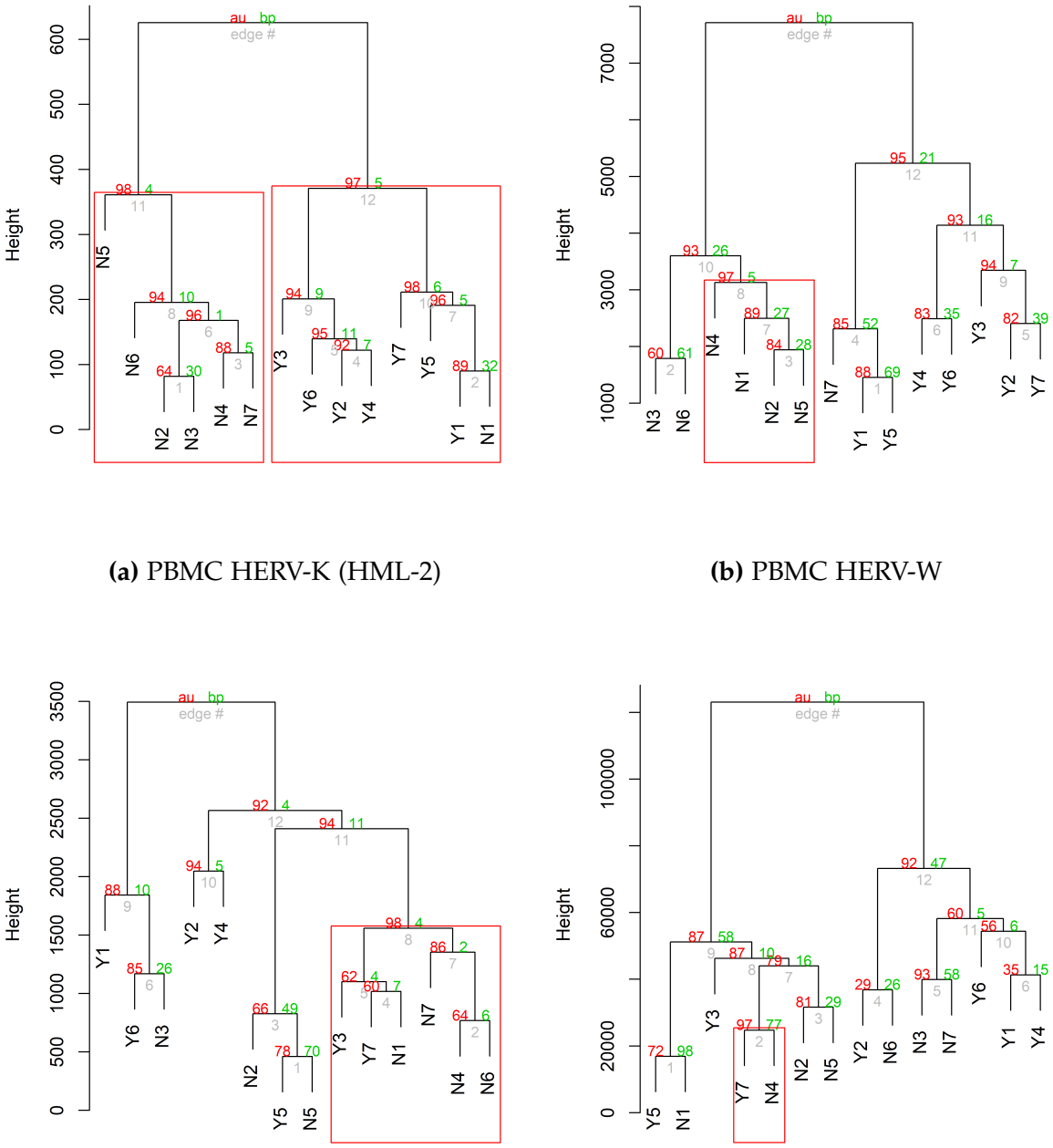
than with HERV-K (HML-2). Clustering of HERV-W is in contrast with PCA results where no discernible clusters were formed.

Clustering of plasma samples did not result in age-associated clusters for either provirus family. It is possible that individual differences in expression in plasma overshadow age group ones. It has been reported that EV concentration and content varies from person to person, strongly enough to have personal EV profiles [13]. This may interfere with the results. A greater number of samples could offer evidence that is more robust. Another possibility, though rather unfeasible, could be longitudinal studies. The issue of personal EV profiles could be avoided by comparing EV samples from the same individuals separated by date of sample collection.

These clustering results can be seen to indicate that there truly are age-associated expression profiles for HERVs. That the differences seen between age groups are not just background noise in the data, but rather similarities in proviral expression within those age groups.

Overall, the results on age group comparisons are not as clear-cut as was expected. No universal upregulation of proviral elements was observed, though clustering results do indicate differing expression patterns between young controls and nonagenarians. It seems that the associations between proviruses and aging are more intricate and may be restricted to certain provirus families, subfamilies or even just individual proviruses.

The question of the underlying mechanism then arises. What could cause these observed differences in proviral expression patterns? One potential explanation is age-related changes in epigenetic silencing. Retrotransposons in general, which HERVs are part of, have for a long time been hypothesized to be kept under control with epigenetic silencing, mainly through methylation [61]. ERVs have been reported to become progressively demethylated and de-silenced in aging mice [37, 7]. This same process could take place in humans, explaining the differential expression between the age groups.



(a) PBMC HERV-K (HML-2)

(b) PBMC HERV-W

(c) Plasma HERV-K (HML-2)

(d) Plasma HERV-W

Figure 6.7: Hierarchical clustering of PBMC samples based on normalized HERV-K and HERV-W read counts, using Spearman correlation distance metric and Ward's minimum variance agglomeration method. The height separating clusters indicates proportional dissimilarity between clusters. The red squares indicate clusters that were deemed statistically significant through bootstrap resampling (p-value < 0.05). AU p-value indicates the bias corrected percentage of dendrogram variants where the cluster was present. Samples from young controls are denoted by the letter Y, and nonagenarians by the letter N.

6.3 Origin of the extracellular vesicles

It was concluded that no definitive conclusions can be made about the origin of the EVs based on only this data, though it can be hypothesized on. The cargo of extracellular vesicles is characteristic of both the type of EV in question as well as the cell of origin [39]. Thus, it was thought that it might be possible to state whether the exosomes are being released from PBMCs or not based on similar expression profiles.

However, one cannot state this purely based on correlations from this sequencing data, as there are some confounding factors. The heterogeneity of the cell types contained in PBMCs is one such factor, as are the types and origins of EVs that can be present in the plasma samples. EVs are thought to be released into circulation by a variety of cell types. These include, but are not limited to, platelets, erythrocytes, neurons, adipocytes, and endothelial cells [13]. Therefore it is very difficult to conclude where exactly the EVs in our samples are from, or if there even is a single distinct source or instead a mixture of sources. Even without these challenges, similarity of expression profiles alone would arguably not be conclusive evidence. As such, the EVs from the plasma samples cannot be stated with confidence to have been secreted from the PBMCs.

One hypothesis on EV origin was revealed by the provirus 19q13.12a. HERV-K (HML-2) provirus 19q13.12a was found to have very high expression in plasma, yet unremarkable expression in PBMCs. The exact location of 19q13.12a is on chromosome 19, at 36063207–36067434, on the minus strand. This provirus has no known aliases. It is instead only referred to by the inhabited locus. 19q13.12a has not been found to contain an open reading frame, thus would not be translated and subsequently is less likely to have biological effects through actual products [25], yet could have biological effects through other means.

Integrative Genomics Viewer [41] was used to investigate the genomic surroundings of 19q13.12a. A view of the IGV interface is shown in figure 6.8 focused on 19q13.12a. No genes or other genomic entities overlap with 19q13.12a in our data. The nearest neighbor of the provirus is the gene ATP4A, located upstream. The database GTEx (Genotype-Tissue Expression) [18] was queried on the gene, which revealed that curiously ATP4A has been reported to only be expressed in gastric parietal cells, located in the stomach (figure 6.9). ATP4A enzyme is needed for gastric acid secretion, and additionally acts as a proton pump, catalyzing the hydrolysis of ATP [36]. If ATP4A is only expressed in gastric parietal cells, then that genomic location should be silenced in other cell types. The high expression of 19q13.12a would indicate that the region cannot be silenced in the plasma sample derived mRNA that

has been analyzed in this work. The mRNA could be delivered from those parietal cells through extracellular vesicles in the bloodstream. However, it is unclear why extracellular vesicles from those cells would be found in the blood stream in the first place.

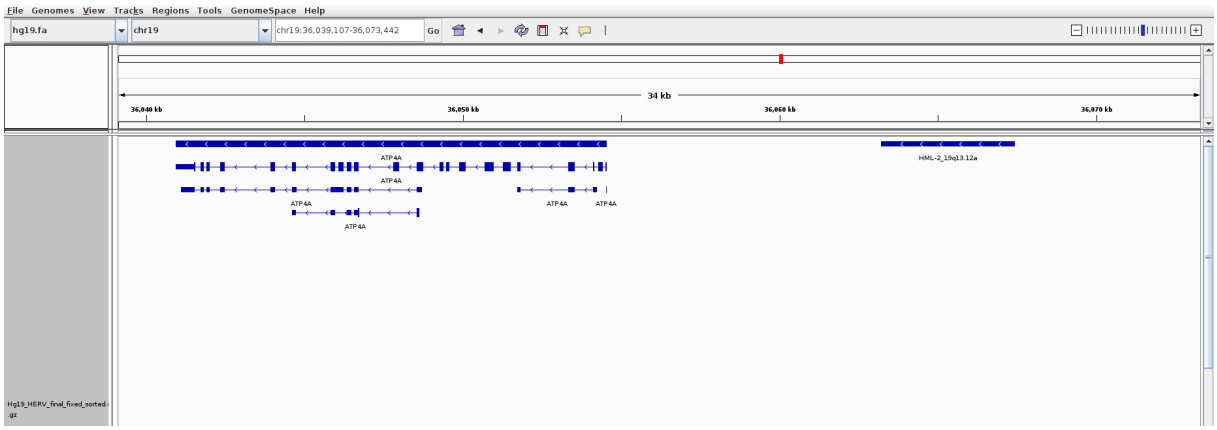


Figure 6.8: Integrative Genomics Viewer showing the genomic location of 19q13.12a and a nearby gene ATP4A.

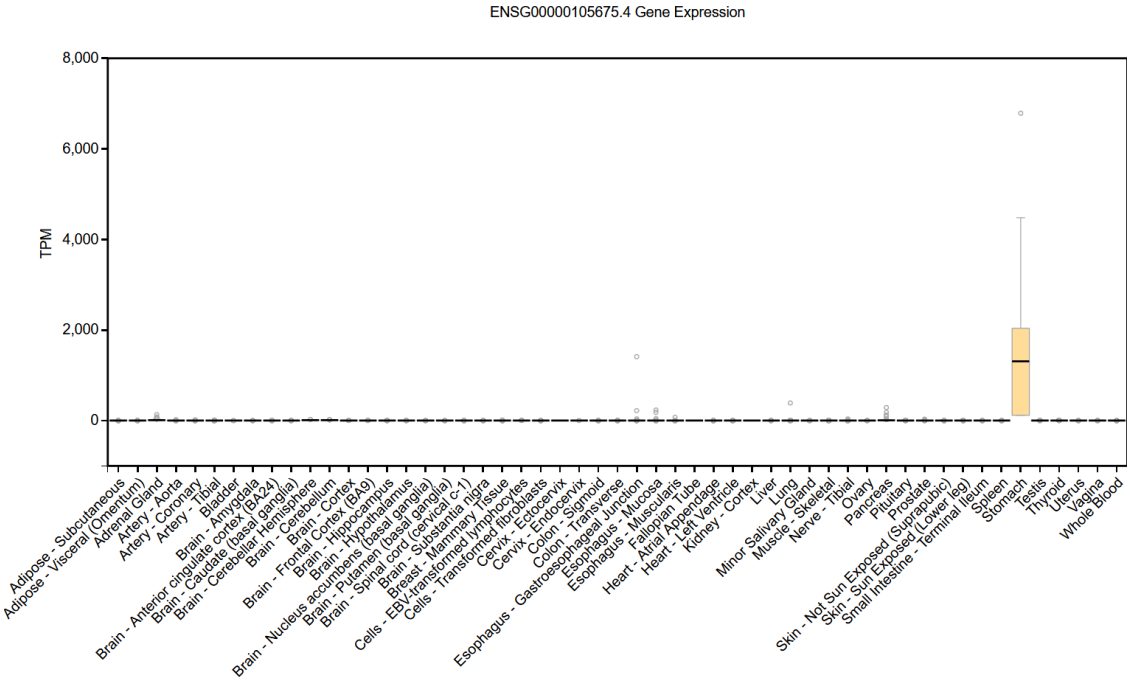


Figure 6.9: View from GTEx database on the tissues in which the gene ATP4A has been reported to be expressed.

6.4 Inferred biological effects

The provirus 19q13.12a had very high expression in plasma, while only having low expression in PBMCs. Due to this rather extreme difference in expression amounts,

GSEA (Gene-Set Enrichment Analysis) was done to identify potential biological effects of 19q13.12a through known functions of genes that have correlating expression with the provirus. Results are displayed in table 6.5. The most significantly enriched GO (Gene Ontology) terms appear very general, and no definitive function for the provirus could be established from them.

Table 6.5: Top 20 results from GSEA of genes correlating with the expression of the provirus 19q13.12a.

	ID	p-value	enrichment score	description
1	GO:0034655	0.0013	0.29	nucleobase-containing compound catabolic process
2	GO:0022613	0.0013	0.44	ribonucleoprotein complex biogenesis
3	GO:0006397	0.0013	0.27	mRNA processing
4	GO:0008380	0.0014	0.30	RNA splicing
5	GO:0034470	0.0014	0.48	ncRNA processing
6	GO:0006401	0.0014	0.33	RNA catabolic process
7	GO:0006402	0.0014	0.35	mRNA catabolic process
8	GO:0000375	0.0014	0.29	RNA splicing, via transesterification reactions
9	GO:0042254	0.0014	0.53	ribosome biogenesis
10	GO:0000377	0.0014	0.30	RNA splicing, via transesterification reactions
11	GO:0000398	0.0014	0.30	mRNA splicing, via spliceosome
12	GO:0090150	0.0014	0.33	establishment of protein localization to membrane
13	GO:0016072	0.0014	0.53	rRNA metabolic process
14	GO:0006364	0.0015	0.55	rRNA processing
15	GO:0071826	0.0015	0.33	ribonucleoprotein complex subunit organization
16	GO:0006403	0.0015	0.29	RNA localization
17	GO:0006399	0.0015	0.35	tRNA metabolic process
18	GO:0000956	0.0015	0.49	nuclear-transcribed mRNA catabolic process
19	GO:0022618	0.0015	0.34	ribonucleoprotein complex assembly
20	GO:0006413	0.0015	0.51	translational initiation

7. Conclusion

The results of this study indicate that some HERV proviruses do have measurable expression in both PBMCs and in EVs, that there are differences in the expression of proviruses between the young and the old, and the measured differences in proviral expression between PBMCs and EVs does support the hypothesis of selectively loaded proviral mRNA cargo of EVs. Age-associated differences were found in the expression of some individual proviruses as well as in proviral expression patterns across samples. These changes in expression could be due to age-related changes in epigenetic silencing. It was not possible to determine with confidence the origin of the studied EVs, nor the biological significance of the abnormally expressed provirus 19q13.12a.

To the best of our knowledge, such a study of proviral expression comparing age groups and using RNA-seq has not been done before. PCR has mainly been used in the past to compare proviral expression [5], yet it is limited in precision and scope, and cannot be used to study the expression of all individual provirus loci across the transcriptome. Use of RNA-seq in this manner shows promise, especially if used with a larger number of samples.

Main limitations of the study approach were identified, such as the small sample size, the heterogeneity of cells contained in PBMC samples, and the potential variation, in terms of type and origin, of the studied EVs. It then becomes difficult to evaluate correlations between contexts, as the cause could be differences in cell proportions, among others. The sample size was enough for an exploratory study such as this, investigating potential avenues of research, yet a greater dataset is needed to support any definitive conclusions.

Perhaps the most important results of this work are not about HERV biology, but rather about how HERVs should be studied in following work. This project has clarified the workflow that we will employ and tools that we should use in the study of HERVs. How these conclusions will guide future studies will be discussed in the next chapter.

8. Future perspectives

The mentioned limitations of this study could be avoided in future studies. The relatively small sample size can in some contexts be avoided by using raw sequencing data from external databases. Gene expression has already been extensively studied with RNA-seq, and it is becoming common practice to release the used raw sequencing data for reproducibility. Data that can be used to study gene expression can likewise be used to study proviral expression. Thus, there are many high-quality datasets available that have been thoroughly investigated in terms of gene expression, yet not in proviral expression. From external databases it could also be possible to find single-cell sequencing data, to avoid the aforementioned problem of multiple cell types contained in PBMCs. Exosome isolated data could possibly also be found that would similarly narrow down the types of EVs in the data. Longer read counts could allow for greater certainty in mapping of proviral transcripts to their originating loci. Studies using additional data might not yield any new insights, but could offer conclusions that are more concrete.

However, there are limitations to using online databases. Concerning the research topic of the biology of aging, there are only few datasets available. Data on nonagenarians is especially rare. High quality sequencing data isolated from EVs may similarly be hard to obtain, though the study of exosomes has been quickly rising in popularity. Therefore, it may be that only some of our research interests can be satisfied with online databases. We are planning to face these difficulties by both expanding our own datasets with data that would otherwise be hard to find, while using external datasets when possible.

We have in this work only investigated the proviruses of HERV-K (HML-2) and HERV-W families. While these two families are some of the most active and therefore of potential biological significance, there are many other families that would be of interest to study. For this task there is for example the online resource HERVd, which is a database on proviruses [38]. It contains information on the location and intactness of 519 060 endogenous retroviral entities. As in the quantification of gene expression, the only thing needed to study the expression of proviruses are their genomic loci.

We are additionally planning to use more up-to-date data analysis pipelines in further studies. Newer, possibly more effective tools such as STAR, Kallisto and DESeq2 could be used instead of Tophat and Cufflinks.

For now, the study of HERVs is the study of correlations. As stated in a review of HERVs and disease [10], more needs to be known about HERV biology for their potential pathology to be understood, and that enough is known for a search for corre-

lation to human disease to be called for. For example the correlations between HERVs and conditions or between HERVs and genes. New technology or another form of a paradigm shift would be required for the nature of HERV research to change. However, even correlation studies could be done better. True leaps in understanding of HERVs will be through utilization of multiple avenues of analysis. Minimizing any issues that could compromise the results while maximizing the strength of the predictions is what should next be done in this research field.

References

[1] Alzheimer's Society. *The MMSE test: Mini Mental State Examination*. URL: https://www.alzheimers.org.uk/info/20071/diagnosis/97/the_mmse_test.

[2] Amaratunga D, Cabrera J, and Kovtun V. "Microarray learning with ABC". In: *Biostatistics* 9.1 (2008), pp. 128–136. DOI: 10.1093/biostatistics/kxm017. eprint: /oup/backfile/content_public/journal/biostatistics/9/1/10.1093_biostatistics_kxm017/1/kxm017.pdf.

[3] Anders S and Huber W. "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10 (2010), R106. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106.

[4] Anderson M, Kashanchi F, and Jacobson S. "Role of Exosomes in Human Retroviral Mediated Disorders". In: *Journal of neuroimmune pharmacology : the official journal of the Society on NeuroImmune Pharmacology* 13.3 (2018), 279—291. ISSN: 1557-1890. DOI: 10.1007/s11481-018-9784-7.

[5] Balestrieri E et al. "Transcriptional Activity of Human Endogenous Retroviruses in Human". In: *BioMed Research International* (2015). DOI: 10.1155/2015/164529.

[6] Baltimore D. "Expression of animal virus genomes." In: *Bacteriol Rev* 35.3 (1971). 4329869[pmid], pp. 235–241. ISSN: 0005-3678.

[7] Barbot W et al. "Epigenetic regulation of an IAP retrotransposon in the aging mouse: progressive demethylation and de-silencing of the element by its repetitive induction". In: *Nucleic Acids Res* 30.11 (2002). 12034823[pmid], pp. 2365–2373. ISSN: 1362-4962.

[8] Baust C et al. "HERV-K-T47D-Related Long Terminal Repeats Mediate Polyadenylation of Cellular Transcripts". In: *Genomics* 66.1 (2000), pp. 98 –103. ISSN: 0888-7543. DOI: https://doi.org/10.1006/geno.2000.6175.

[9] Benjamini Y and Hochberg Y. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 00359246.

[10] Blomberg J, Ushameckis D, and Jern P. "Evolutionary Aspects of Human Endogenous Retroviral Sequences (HERVs) and Disease". In: *Madame Curie Bioscience Database [Internet]. Austin (TX): Landes Bioscience.* (2000-2013).

[11] De Cecco M et al. "Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues". In: *Aging (Albany NY)* 5.12 (2013). 24323947[pmid], pp. 867–883. ISSN: 1945-4589.

[12] D'haeseleer P. "How does gene expression clustering work?" In: *Nature Biotechnology* 23 (2005), 1499 EP –.

[13] Eitan E et al. "Age-Related Changes in Plasma Extracellular Vesicle Characteristics and Internalization by Leukocytes". In: *Scientific Reports* 7.1 (2017), p. 1342. ISSN: 2045-2322. DOI: 10.1038/s41598-017-01386-z.

[14] Escalera-Zamudio M and Greenwood AD. "On the classification and evolution of endogenous retrovirus: human endogenous retroviruses may not be 'human' after all". In: *APMIS* 124.1-2 (), pp. 44–51. DOI: 10.1111/apm.12489. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/apm.12489.

[15] Gardner MB, Kozak C, and O'Brien S. "The Lake Casitas wild mouse: evolving genetic resistance to retroviral disease." In: *Trends Genet.* 7.1 (1991), pp. 22–7.

[16] Gonzalez-Cao M et al. "Human endogenous retroviruses and cancer". In: *Cancer Biol Med* 13.4 (2016). cbm-13-4-483[PII], pp. 483–488. ISSN: 2095-3941.

[17] Grandi N et al. "Contribution of type W human endogenous retroviruses to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes". In: *Retrovirology* 13.1 (2016), p. 67. ISSN: 1742-4690. DOI: 10.1186/s12977-016-0301-x.

[18] GTEx Consortium. "The Genotype-Tissue Expression (GTEx) project". In: *Nat Genet* 45.6 (2013). 23715323[pmid], pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653.

[19] Haase K, Mösch A, and Frishman D. "Differential expression analysis of human endogenous retroviruses based on ENCODE RNA-seq data". In: *BMC Med Genomics* 8 (2015). 146[PII], p. 71. ISSN: 1755-8794. DOI: 10.1186/s12920-015-0146-5.

[20] Hand DJ and Heard NA. "Finding Groups in Gene Expression Data". In: *J Biomed Biotechnol* 2005.2 (2005). 16046827[pmid], pp. 215–225. ISSN: 1110-7243. DOI: 10.1155/JBB.2005.215.

[21] Hayward A. "Origin of the retroviruses: when, where, and how?" In: *Current Opinion in Virology* 25 (2017). Animal models for viral diseases • Paleovirology, pp. 23 –27. ISSN: 1879-6257. DOI: https://doi.org/10.1016/j.coviro.2017.06.006.

[22] Hohn O, Hanke K, and Bannert N. "HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease". In: *Front Oncol* 3 (2013). 24066280[pmid], p. 246. ISSN: 2234-943X. DOI: 10.3389/fonc.2013.00246.

[23] Jintaridth P and Mutirangura A. "Distinctive patterns of age-dependent hypomethylation in interspersed repetitive sequences". In: *Physiological Genomics* 41.2 (2010). PMID: 20145203, pp. 194–200. DOI: 10.1152/physiolgenomics.00146.2009. eprint: https://doi.org/10.1152/physiolgenomics.00146.2009.

[24] Jylhävä J, Pedersen NL, and Hägg S. "Biological Age Predictors". In: *EBioMedicine* 21 (2017), pp. 29 –36. ISSN: 2352-3964. DOI: https://doi.org/10.1016/j.ebiom.2017.03.046.

[25] Karamitros T et al. "A contaminant-free assessment of Endogenous Retroviral RNA in human plasma". In: *Scientific Reports* 6 (2016). Article, 33598 EP –.

[26] Khodosevich K, Lebedev Y, and Sverdlov E. "Endogenous Retroviruses and Human Evolution". In: *Comp Funct Genomics* 3.6 (2002). 18629260[pmid], pp. 494–498. ISSN: 1531-6912. DOI: 10.1002/cfg.216.

[27] Kotlyar M et al. “Spearman Correlation Identifies Statistically Significant Gene Expression Clusters in Spinal Cord Development and Injury”. In: *Neurochemical Research* 27.10 (2002), pp. 1133–1140. ISSN: 1573-6903. DOI: 10.1023/A:1020969208033.

[28] Leale G et al. “Inferring Unknown Biological Function by Integration of GO Annotations and Gene Expression Data”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15.1 (2018), pp. 168–180. ISSN: 1545-5963. DOI: 10.1109/TCBB.2016.2615960.

[29] Li M et al. “Analysis of the RNA content of the exosomes derived from blood serum and urine and its potential as biomarkers”. In: *Philos Trans R Soc Lond B Biol Sci* 369.1652 (2014). rstb20130502[PII], p. 20130502. ISSN: 0962-8436. DOI: 10.1098/rstb.2013.0502.

[30] Macfarlane CM and Badge RM. “Genome-wide amplification of proviral sequences reveals new polymorphic HERV-K(HML-2) proviruses in humans and chimpanzees that are absent from genome assemblies”. In: *Retrovirology* 12 (2015). 162[PII], p. 35. ISSN: 1742-4690. DOI: 10.1186/s12977-015-0162-8.

[31] Marttila S. “Ageing-associated changes in gene expression and DNA methylation with implications for intergenerational epigenetic inheritance”. PhD dissertation. University of Tampere, 2016.

[32] Marttila S et al. “Human endogenous retrovirus HERV-K(HML-2) env expression is not associated with markers of immunosenescence”. In: *Experimental Gerontology* 97 (2017), pp. 60 –63. ISSN: 0531-5565. DOI: https://doi.org/10.1016/j.exger.2017.07.019.

[33] McDonald J. *Handbook of Biological Statistics (3rd ed.)* Sparky House Publishing, 2014, pp. 254–260.

[34] Nelson PN et al. “Demystified. Human endogenous retroviruses”. In: *Mol Pathol* 56.1 (2003). 0560011[PII], pp. 11–18. ISSN: 1366-8714.

[35] Nevalainen T et al. “Obesity accelerates epigenetic aging in middle-aged but not in elderly individuals”. In: *Clinical Epigenetics* 9.1 (2017), p. 20. ISSN: 1868-7083. DOI: 10.1186/s13148-016-0301-7.

[36] O’Leary NA et al. “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: *Nucleic Acids Res* 44.D1 (2016). 26553804[pmid], pp. D733–D745. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

[37] Ono T et al. “Endogenous virus genomes become hypomethylated tissue—Specifically during aging process of C57BL mice”. In: *Mechanisms of Ageing and Development* 50.1 (1989), pp. 27 –36. ISSN: 0047-6374. DOI: https://doi.org/10.1016/0047-6374(89)90056-0.

[38] Paces J, Pavlícek A, and Paces V. “HERVd: database of human endogenous retroviruses”. In: *Nucleic Acids Res* 30.1 (2002). gkf077[PII], pp. 205–206. ISSN: 0305-1048.

[39] Prendergast EN et al. “Optimizing exosomal RNA isolation for RNA-Seq analyses of archival sera specimens”. In: *PLOS ONE* 13.5 (May 2018), pp. 1–14. DOI: 10.1371/journal.pone.0196913.

[40] Qin C et al. "Intracisternal A particle genes: Distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer". In: *Molecular Carcinogenesis* 49.1 (), pp. 54–67. DOI: 10.1002/mc.20576. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mc.20576.

[41] Robinson JT et al. "Integrative Genomics Viewer". In: *Nature Biotechnology* 29 (2011), 24–26.

[42] Roy S, Hochberg FH, and Jones PS. "Extracellular vesicles: the growth as diagnostics and therapeutics; a survey". In: *J Extracell Vesicles* 7.1 (2018). 1438720[PII], p. 1438720. ISSN: 2001-3078. DOI: 10.1080/20013078.2018.1438720.

[43] Schurch NJ et al. "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" In: *RNA* 22.6 (2016). RA[PII], pp. 839–851. ISSN: 1355-8382. DOI: 10.1261/rna.053959.115.

[44] ScienceDirect Topics. *Barthel scale - an overview*. URL: https://www.sciencedirect.com/topics/medicine-and-dentistry/barthel-scale.

[45] Silver N et al. "Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR". In: *BMC Mol Biol* 7 (2006). 1471-2199-7-33[PII], pp. 33–33. ISSN: 1471-2199. DOI: 10.1186/1471-2199-7-33.

[46] Sokol M, Jessen KM, and Pedersen FS. "Utility of next-generation RNA-sequencing in identifying chimeric transcription involving human endogenous retroviruses". In: *APMIS* 124.1-2 (), pp. 127–139. DOI: 10.1111/apm.12477. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/apm.12477.

[47] Subramanian RP et al. "Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses". In: *Retrovirology* 8 (2011). 1742-4690-8-90[PII], pp. 90–90. ISSN: 1742-4690. DOI: 10.1186/1742-4690-8-90.

[48] Suzuki R and Shimodaira H. "Pvclust: an R package for assessing the uncertainty in hierarchical clustering". In: *Bioinformatics* 22.12 (2006), pp. 1540–1542. DOI: 10.1093/bioinformatics/btl117.

[49] The Gene Ontology Consortium. "Expansion of the Gene Ontology knowledgebase and resources". In: *Nucleic Acids Res* 45.Database issue (2017). 27899567[pmid], pp. D331–D338. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1108.

[50] The Gene Ontology Consortium. "Gene Ontology: tool for the unification of biology". In: *Nat Genet* 25.1 (2000). 10802651[pmid], pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556.

[51] Ting CN et al. "Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene." In: *Genes & Development* 6.8 (1992), pp. 1457–1465. DOI: 10.1101/gad.6.8.1457. eprint: http://genesdev.cshlp.org/content/6/8/1457.full.pdf+html.

[52] Trapnell C, Pachter L, and Salzberg SL. "TopHat: discovering splice junctions with RNA-Seq". In: *Bioinformatics* 25.9 (2009). 19289445[pmid], pp. 1105–1111. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp120.

[53] Trapnell C et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq". In: *Nat Biotechnol* 31.1 (2013). 23222703[pmid], pp. 46–53. ISSN: 1546-1696. DOI: 10.1038/nbt.2450.

[54] Trapnell C et al. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks”. In: *Nat Protoc* 7.3 (2012). 22383036[pmid], pp. 562–578. ISSN: 1750-2799. DOI: 10.1038/nprot.2012.016.

[55] Urbanelli L et al. “Extracellular Vesicles as New Players in Cellular Senescence”. In: *Int J Mol Sci* 17.9 (2016). Ed. by Drummen G. ijms-17-01408[PII], p. 1408. ISSN: 1422-0067. DOI: 10.3390/ijms17091408.

[56] Vargiu L et al. “Classification and characterization of human endogenous retroviruses; mosaic forms are common”. In: *Retrovirology* 13.1 (2016), p. 7. ISSN: 1742-4690. DOI: 10.1186/s12977-015-0232-y.

[57] Venkatesan A and Johnson RT. “Chapter 7 - Infections and multiple sclerosis”. In: *Multiple Sclerosis and Related Disorders*. Ed. by Goodin DS. Vol. 122. Handbook of Clinical Neurology. Elsevier, 2014, pp. 151 –171. DOI: https://doi.org/10.1016/B978-0-444-52001-2.00007-8.

[58] Wada Y et al. “Retroviral gene expression as a possible biomarker of aging”. In: *International Archives of Occupational and Environmental Health* 65.1 (1993), S235–S240. ISSN: 1432-1246. DOI: 10.1007/BF00381349.

[59] Wang-Johanning F et al. “Human Endogenous Retrovirus Type K Antibodies and mRNA as Serum Biomarkers of Early-Stage Breast Cancer”. In: *Int J Cancer* 134.3 (2014). 23873154[pmid], pp. 587–595. ISSN: 0020-7136. DOI: 10.1002/ijc.28389.

[60] Weiss RA. “The discovery of endogenous retroviruses”. In: *Retrovirology* 3.1 (2006), p. 67. ISSN: 1742-4690. DOI: 10.1186/1742-4690-3-67.

[61] Yoder JA, Walsh CP, and Bestor TH. “Cytosine methylation and the ecology of intragenomic parasites”. In: *Trends in Genetics* 13.8 (1997). Epigenetics, pp. 335 –340. ISSN: 0168-9525. DOI: https://doi.org/10.1016/S0168-9525(97)01181-5.

[62] Young GR, Stoye JP, and Kassiotis G. “Are human endogenous retroviruses pathogenic?: An approach to testing the hypothesis”. In: *Bioessays* 35.9 (2013). 23864388[pmid], pp. 794–803. ISSN: 0265-9247. DOI: 10.1002/bies.201300049.

[63] Zappulli V et al. “Extracellular vesicles and intercellular communication within the nervous system”. In: *J Clin Invest* 126.4 (2016). 81134[PII], pp. 1198–1207. ISSN: 0021-9738. DOI: 10.1172/JCI81134.

[64] Zhang R et al. “Interaction of Epstein-Barr virus genes with human gastric carcinoma transcriptome”. In: *Oncotarget* 8.24 (2017). 28415594[pmid], pp. 38399–38412. ISSN: 1949-2553.